## Detecting Copy Number Variations from Next-Generation Sequencing Data via (2) a Bayesian Procedure

Yu-Chung Wei and Guan-Hua Huang

National Chiao Tung University

## Detecting Copy Number Variations from Next－Generation Sequencing Data via a Bayesian Procedure

Yu－Chung Wei and Guan－Hua Huang
国立交通大業
National Chiao Tung University

# - Copy Number 

 Variations- Sequencing Read Depths
- Bayesian

Procedure

- Model
-RJMCMC
- NTUH data


- NTUH data


# - Copy Number 

 Variations- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data


| T | T | C | G | A | A | $\ldots$ | C | G | T | A | A | T | C | G | T | A | A | T |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| A | A | G | C | T | T | $\ldots$ | C | G | C | A | T | T | A | G | C | A | T | T | $>1000$ bases

## $\mathrm{CN}=2$

## $\mathrm{CN}=4$

- Copy Number Variations
- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data


## Copy number variations (CNVs) and human genetic diseases

- CNVs account for roughly $12 \%$ of the human genome
- Down syndrome: a genetic disorder caused by the presence of a third copy of chromosome 21
- Mental disorders, including autism, schizophrenia: about $1 \%$ with rare DNA deletions in chromosome 15q13.3, 16p11.2, or 1q21.1
Breast cancer: 20-30\% with HER-2 gene amplification and over-expression
- Copy Number Variations


## Detecting CNVs

- Sequencing Read Depths Array-based comparative genome hybridization (array-CGH)
- Bayesian Procedure -Model -RJMCMC
- NTUH data
a Reference DNA


Test DNA
Hybridize to arrays


Block repeats with COT-1 DNA

Detect and quantify signals (Cy3:Cy5)

Spurious signal


- Copy Number Variations


## Detecting CNVs

## Spotted oligonucleotides on Affymetrix SNP

 arrays- Bayesian Procedure -Model -RJMCMC
- NTUH data

- Copy Number Variations


## Limitations of hybridizationbased microarray approaches

- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data
- Hybridization-based microarray approaches: array-CGH and SNP arrays
- Microarrays are limited to
- detecting copy-number differences of sequences present in the reference assembly used to design the probes,
- provide no information on the location of duplicated copies,
- are generally unable to resolve breakpoints at the single-base-pair level
- Copy Number Variations
- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data


## Sequencing-based computational approaches

- The advent of next-generation sequencing (NGS) technologies promises to revolutionize copy number variation (CNV).
- NGS approaches can map CNVs with much greater accuracy than hybridization-based microarray approaches.
However, NGS approaches present substantial computational and bioinformatics challenges.
- Copy Number Variations
- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data


## Sequencing-based computational approaches

- There are four general types of NGS strategy, all of which focus on mapping sequence reads to the reference genome and subsequently identifying CNVs:
- read-pair (paired-end reads),
- read-depth,
- split-read,
- sequence assembly.
- Copy Number Variations


## $A$ A C T T $\cdots \cdots$ G C A T A G C A T T A

O Sequencing Read Depths


- Bayesian

Procedure
-Model
-RJMCMC

$$
\begin{array}{|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|c|}
\hline \mathrm{T} & \mathrm{~T} & \mathrm{C} & \mathrm{G} & \mathrm{~A} & \mathrm{~A} & \ldots . . & \mathrm{C} & \mathrm{G} & \mathrm{~T} & \mathrm{~A} & \mathrm{~A} & \mathrm{~T} & \mathrm{C} & \mathrm{G} & \mathrm{~T} & \mathrm{~A} & \mathrm{~A} & \mathrm{~T} \\
\hline \mathrm{~A} & \mathrm{~A} & \mathrm{G} & \mathrm{C} & \mathrm{~T} & \mathrm{~T} & \ldots . . & \mathrm{G} & \mathrm{C} & \mathrm{~A} & \mathrm{~T} & \mathrm{~T} & \mathrm{~A} & \mathrm{G} & \mathrm{C} & \mathrm{~A} & \mathrm{~T} & \mathrm{~T} & \mathrm{~A} \\
\hline
\end{array}
$$

$$
\begin{array}{|c|c|c|c|c|c|cc|c|c|c|c|c|c|c|c|c|c|c|}
\hline \mathrm{T} & \mathrm{~T} & \mathrm{C} & \mathrm{G} & \mathrm{~A} & \mathrm{~A} & \ldots . . & \mathrm{C} & \mathrm{G} & \mathrm{~T} & \mathrm{~A} & \mathrm{~A} & \mathrm{~T} & \mathrm{C} & \mathrm{G} & \mathrm{~T} & \mathrm{~A} & \mathrm{~A} & \mathrm{~T} \\
\hline \mathrm{~A} & \mathrm{~A} & \mathrm{G} & \mathrm{C} & \mathrm{~T} & \mathrm{~T} & \ldots . . & \mathrm{G} & \mathrm{C} & \mathrm{~A} & \mathrm{~T} & \mathrm{~T} & \mathrm{~A} & \mathrm{G} & \mathrm{C} & \mathrm{~A} & \mathrm{~T} & \mathrm{~T} & \mathrm{~A} \\
\hline
\end{array}
$$

- NTUH data
- Copy Number Variations

O Sequencing Read Depths

- Bayesian Procedure -Model -RJMCMC
- NTUH data

Aanamm … actamatacanmal



- Copy Number Variations
- Sequencing Read Depths
- Bayesian Procedure -Model $\bullet$ •RJMCMC
- NTUH data




- Copy Number Variations

$$
A, G, C, T, C, A, T, A, G, A
$$

O Sequencing Read Depths

- Bayesian Procedure

$$
A, A, C, T, \quad G, A T, A, C, T, T
$$ -Model -RJMCMC

$$
A, A, C T, T, G, T, A, C, T, A
$$

$A, A, C, T, T$
$G, C, T, T, G, C, T, A$

- NTUH data

$$
A, A_{1} G_{T} T, \quad G, A, T, A, C, A
$$

$$
A, A, C T T, G, A T T A G C A T T A
$$



- Copy Number Variations

O Sequencing Read Depths

- Bayesian Procedure -Model $\bullet$ •RJMCMC
- NTUH data

 $\begin{array}{lll}0.0 & 0.2 & 0.4\end{array}$ density
- Copy Number log RD /window:
- Sequencing

- Bayesian Procedure - Model $\bullet$ •RJMCMC
- NTUH data

$$
\begin{aligned}
& P(\mathbf{C}, \underline{\mathbf{B}} \mathbf{D}) \\
\propto & \left.P(\underset{\sim}{\mathbf{D}} \mid \mathbf{C}, \underline{\mathbf{B}}) \times P\left(\underset{\sim}{\mathbf{C}}, \mid \mathbf{P}_{\underline{\mathbf{B}}}\right)\right) \times P(\underset{\sim}{\mathbf{B}})
\end{aligned}
$$

- Copy Number $\log$ RD $\begin{array}{llllllllll}D_{1} & D_{3} & D_{5} & D_{7} & D_{9} & D_{11} & D_{13} & D_{15} & D_{17} & D_{19} \\ D_{21} & D_{23} & D_{25}\end{array}$ Variations /window: $\begin{array}{llllll}D_{2} & D_{4} & D_{6} & D_{8} & D_{10} & D_{12} \\ D_{14} & D_{16} & D_{18} & D_{20} & D_{22} & D_{24}\end{array}$

- Sequencing Read Depths
- Bayesian Procedure - Model -RJMCMC

| Breakpoint |  |  | ¢-18 |  |
| :---: | :---: | :---: | :---: | :---: |
|  | ...... |  |  |  |
|  |  |  |  |  |
| CN state: | $\mathrm{C}_{1}{ }^{*}=2$ | $C_{2}{ }^{*}=1$ | $\mathrm{C}_{3}{ }^{*}=2$ | $\mathrm{C}_{4}{ }^{*}=3$ |


$w_{01}, w_{02}, \ldots, w_{0 k}$

$$
\frac{W_{01}}{1-W_{02}}, \frac{W_{03}}{1-W_{02}}, \frac{W_{0 K}}{1-W_{02}}
$$

- NTUH data

$$
\begin{aligned}
& P(\mathbf{C}, \underline{\mathbf{B}} \mathbf{D}) \\
\propto & P(\underset{\sim}{\mathbf{D}} \mid \underset{C}{\mathbf{C}}, \underset{\sim}{\mathbf{B}}) \times P(\underset{\sim}{\mathbf{C}} \mid \underset{\sim}{\mathbf{B}}) \times P(\underset{\sim}{\mathbf{B}})
\end{aligned}
$$

- Copy Number $\log$ RD $\begin{array}{llllllllllll}D_{1} & D_{3} & D_{5} & D_{7} & D_{9} & D_{11} & D_{13} & D_{15} & D_{17} & D_{19} & D_{21} & D_{23} \\ D_{25}\end{array}$ /window: $\begin{array}{lllllll}D_{2} & D_{4} & D_{6} & D_{8} & D_{10} & D_{12} & D_{14} \\ D_{16} & D_{18} & D_{20} & D_{22} & D_{24}\end{array}$ Variations
- Sequencing Read Depths $\begin{array}{rllll}\mathbf{B}_{0} \mathbf{B}_{1} & & \mathbf{B}_{8} \mathbf{B}_{9} & \mathbf{B}_{13} & \mathbf{B}_{20} \\ \text { Breakpoint: }\|\| & \cdots \cdots & \|\|\| & \| & \| \\ \mathbf{1} 0 & & \mathbf{0} \mathbf{1} & \mathbf{1} & \mathbf{1}\end{array}$ CN state:

$$
\mathrm{C}_{1}{ }^{*}=2
$$

$$
C_{2}^{*}=1
$$

$$
C_{3}^{*}=2
$$

$$
\mathrm{C}_{4}{ }^{*}=3
$$

- Bayesian Procedure - Model -RJMCMC


## $B_{i} \sim \operatorname{Binomial}(\boldsymbol{\lambda})$

- NTUH data

$$
\begin{aligned}
& \quad P(\underset{\sim}{\mathbf{C}}, \underset{\mathbf{B}}{\mathbf{D}}) \\
& \propto P(\underset{\sim}{\mathbf{D}} \mid \underset{\mathbf{C}}{\mathbf{B}}, \underset{\sim}{\mathbf{B}}) \times P(\underset{\sim}{\mathbf{C} \mid \underset{\sim}{\mathbf{B}}) \times P(\underset{\sim}{\mathbf{B}})}
\end{aligned}
$$

- Copy Number $\log$ RD /window:

$$
\begin{array}{lllllllllll}
D_{1} & D_{3} & D_{5} & D_{7} & D_{9} & D_{11} & D_{13} & D_{15} & D_{17} & D_{19} & D_{21}
\end{array} D_{23} D_{25}
$$ Variations

- Sequencing Read Depths $\begin{array}{ccccc}\mathbf{B}_{0} \mathbf{B}_{1} & & \mathbf{B}_{8} \mathbf{B}_{9} & \mathbf{B}_{13} & \mathbf{B}_{20} \\ \text { Breakpoint: }\|\| & \cdots \cdots & \left\|\|_{1}\right. & \| & \| \\ \mathbf{1} 0 & & \mathbf{0} \mathbf{1} & \mathbf{1} & \mathbf{1}\end{array}$ CN state:
$\mathrm{C}_{1}{ }^{*}=2$
$C_{2}^{*}=1$
$C_{3}^{*}=2$
$C_{4}^{*}=3$
- Bayesian Procedure - Model -RJMCMC
- NTUH data
- Copy Number Variations


## 1. Merge

- Sequencing Read Depths

CN state: $\square$

- Bayesian

Procedure -Model -RJMCMC

- NTUH data

$$
\mathrm{C}_{1}^{*}=2
$$

$$
C_{2}^{*}=3
$$

- Copy Number Variations
- Sequencing Read Depths
- Bayesian

Procedure -Model -RJMCMC

- NTUH data


## 2. Split



- Copy Number Variations


## 3. Trifid

- Bayesian Procedure -Model -RJMCMC
- NTUH data

CN state:

| $\mathrm{C}_{1}^{*}=2$ | $\mathrm{C}_{2}^{*}=1$ | $\mathrm{C}_{3}^{*}=2$ | $\mathrm{C}_{4}^{*}=3$ |
| :--- | :--- | :--- | :--- |



- Copy Number Variations


## 4. Boundary Change

- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data

CN state:


- Copy Number Variations
- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data


## NTUH Department of OB/GYN

- Silver-Russell syndrome: a growth disorder, have a small, triangular face with distinctive facial features
- Illumina/Solexa (NGS technology)
- Targeted exon region (protein coding regions)
- Chromosome 7
- 32387 windows
- Copy Number Variations


## Sample 1:

- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data

$\bullet: C N 1 \quad$ :CN2 •:CN3 •:CN4 •:CN5


# - Copy Number 

 Variations
## Sample 2:

- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data

$\bullet: C N 1$ •:CN2 •:CN3 •:CN4 •:CN5


# - Copy Number 

 Variations
## Sample 1 vs. 2:

- Sequencing Read Depths
- Bayesian Procedure -Model -RJMCMC
- NTUH data

- :Deletion
- :Normal
- :Duplication

